Character Keypoint-based Homography Estimation in Scanned Documents for Efficient Information Extraction

Kushagra Mahajan, Monika Sharma, Lovekesh Vig
TCS Research, New Delhi, India

Introduction

- Homography estimation for aligning test document images (contains user-filled information) with a template document (unfilled document) for efficient information extraction from documents.
- Propose a novel, robust and memory efficient algorithm for keypoint extraction from scanned or camera-captured document images.
- Uses distinct tips in the individual characters of the document.
- Keypoint correspondences between the template and test documents used to estimate homography. The test document is aligned with the template. Fields in the test document are extracted.
- The extracted text (printed and handwritten) is then read using pre-existing works.

Motivation

- **Question.** How can we get the information added by hundreds of users on forms like insurance forms, bank receipts, job application forms etc. through their images in an automated process?

- For facilitating fast retrieval of information, digitization is being used in every aspect of industry.
  - Machinery logs in factories
  - Contracts in government offices.
  - Scanned or camera-captured document images such as bank receipts, insurance claim forms etc.

- Scanned or camera-captured documents have variations in **orientations** and **illumination**, which cause automated information retrieval systems to be prone to errors thereby requiring manual intervention.

- Document alignment **reduces errors** in automated information extraction and also **reduces time and costs** for digitization of scanned documents since no manual intervention is involved.

Related Work

- Why develop specialized techniques for documents? There are already numerous image alignment techniques in the literature.
  - Direct pixel-based: Lucas Kanade's Optical Flow [1], Lucey et al's work [2]
  - Feature-based: SIFT [3], ORB [4] descriptors.

- Some other works that focus explicitly on alignment of documents.
  - Takeda et al [5] uses the centroids of words in the document to compute the features.
  - Block et al [6] exploited structures in the text document like punctuation characters as keypoints for document mosaicing
  - Royer et al [7] explored keypoint selection methods which reduces the number of extracted keypoints for improved document image matching.

- Other relevant works that have been used for image alignment.
  - Rocco et al [8] proposed an end-to-end architecture for learning affine transformations without manual annotation through a siamese architecture.
  - DeTone et al [9] devised a DNN to estimate the displacements between the four corners of the original and perturbed images in a supervised manner, and map it to the corresponding homography matrix.
  - Nguyen et al [10] trains a CNN for unsupervised learning of planar homographies, achieving faster inference and superior performance compared to the supervised counterparts.

Figure 1. Keypoint detection using the SIFT descriptor in straight and rotated images

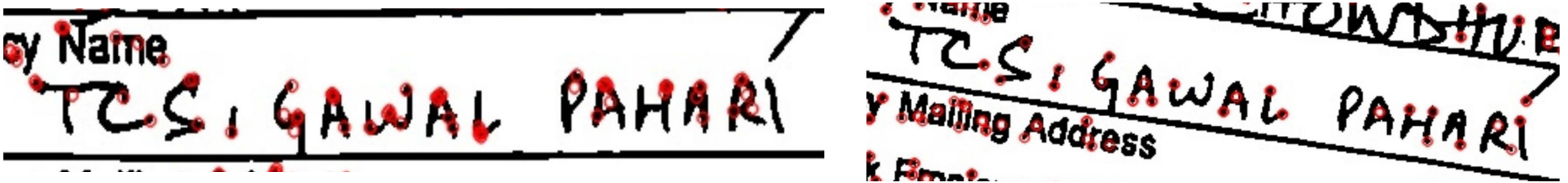


Figure 2. Keypoint detection using the ORB descriptor in straight and rotated images

Contributions

- Propose a novel, fast and memory efficient algorithm for robust **character based unambiguous keypoint detection**, extracted using a standard OCR like Tesseract.
- Demonstrate how existing homography estimation approaches perform poorly when extended to scanned or camera-captured document images. The limitations were analyzed to come up with our methodology.
- We show the effectiveness of our proposed approach using information extraction from two real world anonymized datasets comprised of health insurance claim forms.

Approach

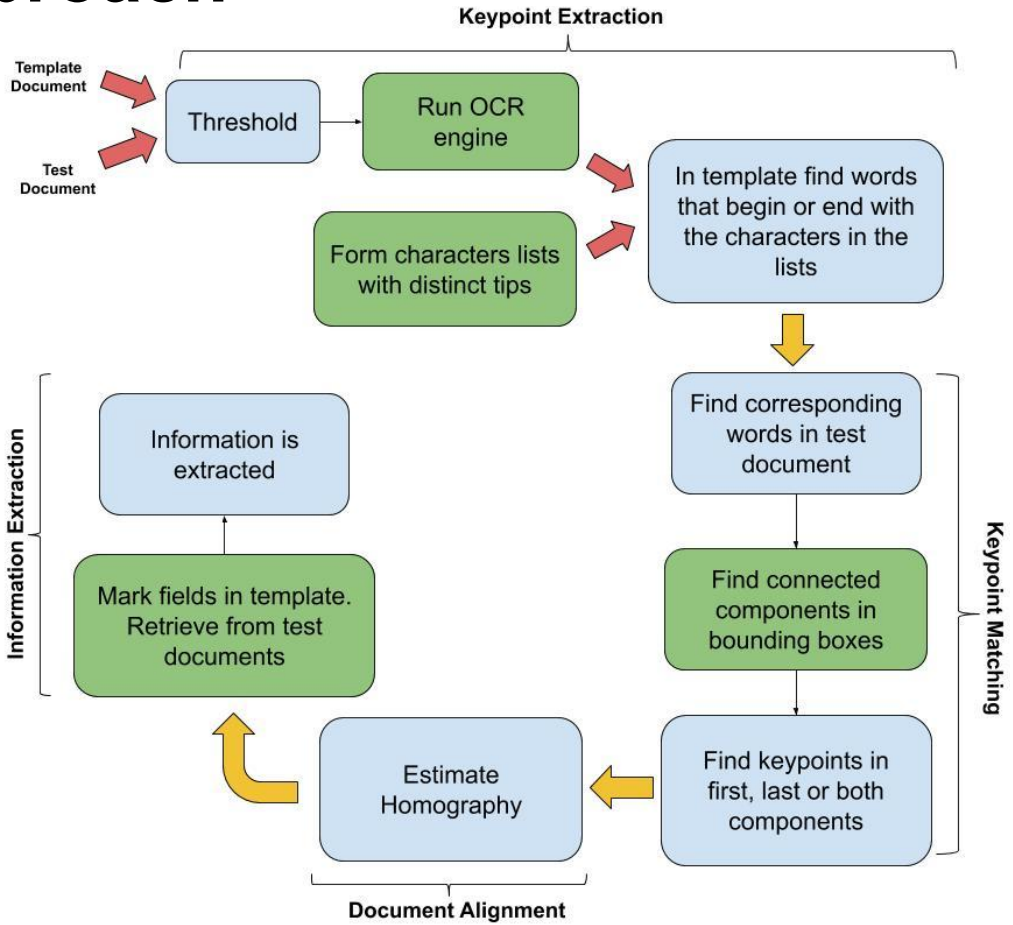


Figure 3. Flowchart showing the entire pipeline for information extraction from scanned or camera-captured document images

Approach (Contd..)

TCS RESEARCH & INNOVATION

- Take the template and test document images. Resize them to 1600px x 2400px. Then, threshold the images. Thresholding allows us to mitigate the impact of illumination variations.
- Form lists of characters with distinct left, right, top and bottom tip points. For example, characters such as 'A', 'V', 'T', 'Y', '4', 'v', 'w' etc. have a distinct left tip, and characters like 'V', 'T', 'L', 'Y', '7', 'r' etc. have a distinct right tip.
- Run an OCR engine like Tesseract to detect words in the template and test documents along the their bounding box coordinates. Look for words beginning or ending with the characters in the 4 lists mentioned in the previous step.
- Select such words in the template document. Use neighbourhood information to find accurate corresponding words in the test document.
- Find connected components in the bounding boxes of the words. Extract keypoints in the first, last or both the components depending on the characters present at these positions in the corresponding words. The keypoints will be the distinct tip points present in these characters.

Approach (Contd..)

- Use the corresponding keypoints obtained to estimate the homography. Then, align the test document image with the template document image.
- The user must mark bounding boxes around the fields to be extracted in the template document image. Corresponding fields are automatically extracted in the test document images.
- The extracted fields could be printed or handwritten text. Train binary classifier to distinguish printed and handwritten text patches.
- Printed text is recognized using an OCR such as Tesseract in our case.
- For handwritten text, we used both the Google Vision API and our own HTR model [13] for recognition. Their accuracies have been compared in tables in the results section later.

Approach (Contd..)

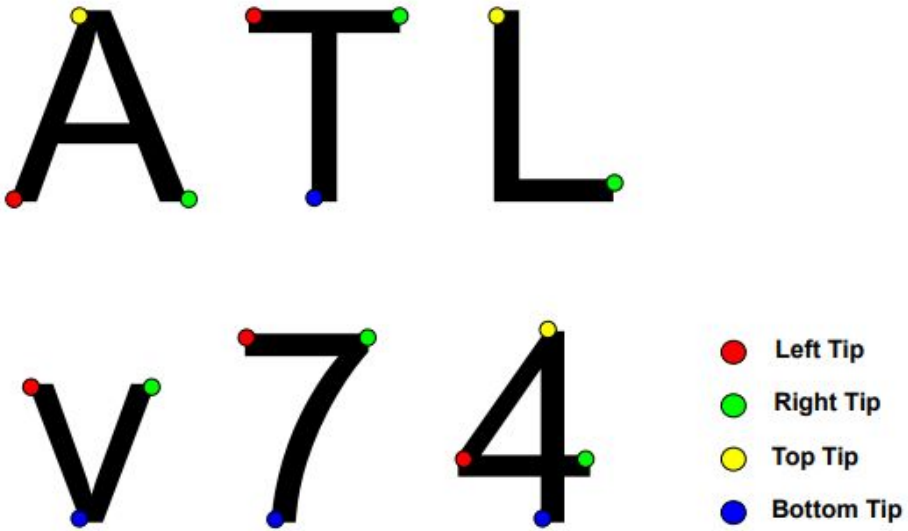
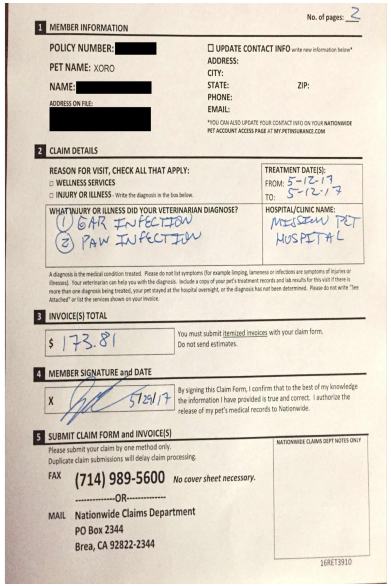


Figure 4. Left, right, top and bottom tips are shown for some of the characters included in the *begCharList*, *endCharList*, *topCharList* and *bottomCharList* respectively.

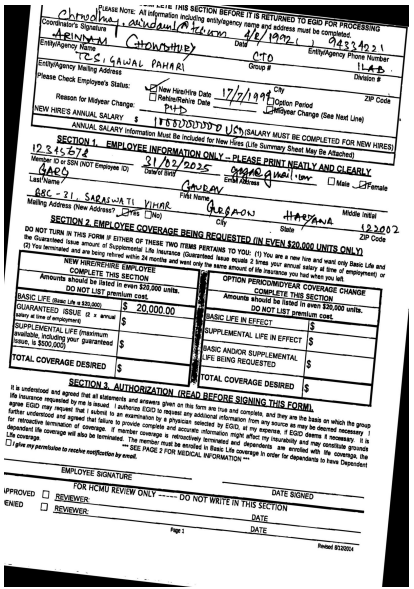
Experimental Dataset

- Evaluated our proposed approach on two real world anonymized document datasets:
  - The first dataset (Insurance1): 15 insurance claim forms and one corresponding empty template form.
  - The second dataset (Insurance2): 15 life insurance application forms and one corresponding empty template form.
- Datasets contain variations in illumination, different backgrounds like wooden table and affine transformed document images.

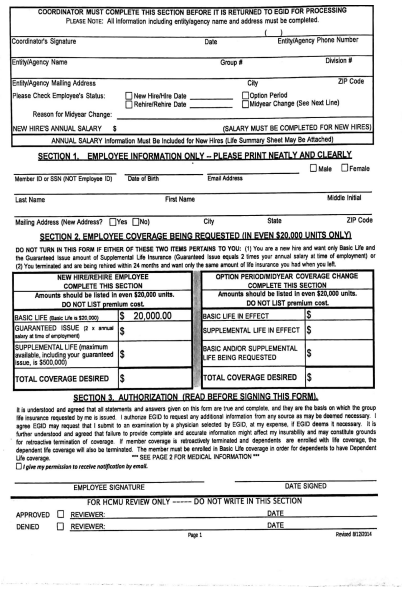


Test Image

Template Image



Test Image

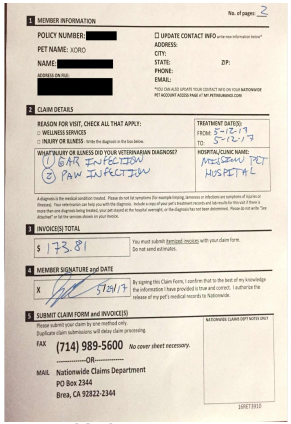


Template Image

Experimental Details

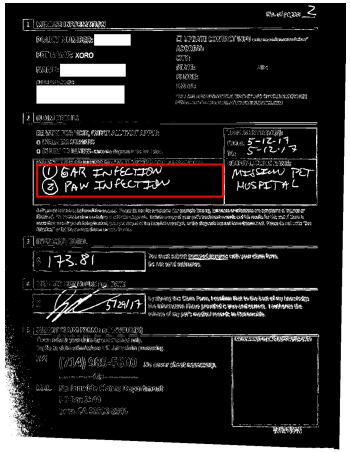
- Alignment is followed by text field retrieval and classification of the text into printed or handwritten. We train a 5-layer CNN.
  - Patches of printed text obtained from text lines detected by CTPN [11] on a separate dataset, and patches of handwritten text obtained from the IAM dataset [12].
  - We obtain a test accuracy of **98.5%** on the combination of the 2 datasets used for experiments.
- Our algorithm is able to handle translations, rotations, and scaling of the test documents.
  - For rotations, the system performance is unaffected for rotations upto ±7 degrees in the x-y plane of the image.
  - Horizontal and vertical translations range in between ±40% of the document width and height respectively.
  - For our datasets, scaling works perfectly when the width and height are varied from 50% to 200% of their original values.

Results

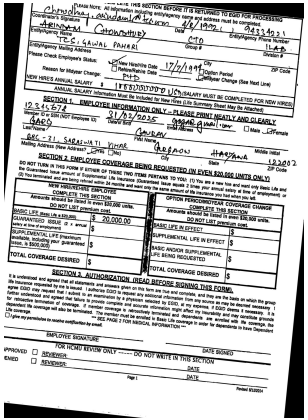


(1a)

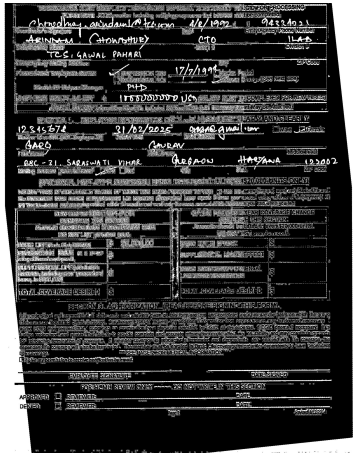
(1b)



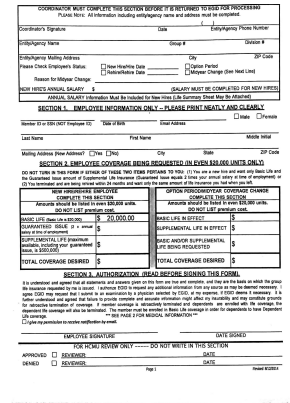
(1c)



(2a)



(2c)



(2b)

Figure 6. Qualitative results for document alignment

Results

CHARACTER RECOGNITION ACCURACY FOR FIELDS IN THE FIRST INSURANCE DATASET. COLUMN (A) GIVES THE ACCURACY OF THE PRINTED TEXT, COLUMN (B) SHOWS THE ACCURACY FOR HANDWRITTEN TEXT TESTED ON THE HTR [19], WHILE COLUMN (C) MENTIONS THE ACCURACY OF HANDWRITTEN TEXT USING THE GOOGLE VISION API.

Field	Tesseract (A)	HTR [19](B)	Vision API (C)
Name	98.6%	88%	92.2%
Pet Name	99.2%	89.5%	92.9%
Address	98.3%	80.4%	85.8%
Hospital	98.7%	77.3%	82.5%
Injury	97.1%	78%	82.6%

Table 1

CHARACTER RECOGNITION ACCURACY FOR FIELDS IN THE SECOND INSURANCE DATASET OF APPLICATION FORMS. COLUMN (A) REPORTS THE ACCURACY FOR HANDWRITTEN TEXT TESTED ON THE HTR MODEL GIVEN BY ARINDAM ET AL., WHILE COLUMN (B) GIVES THE ACCURACY OF HANDWRITTEN TEXT USING THE GOOGLE VISION API. THIS DATASET DOES NOT CONTAIN ADDED TEXT IN PRINTED FORM.

Field	HTR Model [19] (A)	Google Vision API (B)
Agency Name	78.7%	83.5%
Agency Address	78.3%	84.6%
First Name	80.1%	84.5%
Last Name	80.7%	86.7%
Applicant Address	78.4%	82.6%
City	81.9%	93.5%
State	83.2%	89.6%

Table 2

Conclusion

- We proposed a character keypoint-based approach for homography estimation using textual information present in the document to address the problem of image alignment, for scanned or camera-captured textual document images.
- We cannot use the contemporary machine learning and deep learning algorithms since such documents do not have smooth pixel intensity gradients for warp estimation.
- Feature descriptors like SIFT, ORB etc. produce a large number of inconsistent keypoints due to sharp textual edges producing inaccurate keypoint correspondences.
- To address these limitations, we create an automated system which takes an empty template document image and the corresponding filled test document, and aligns the test document with the template for extraction and analysis of textual fields.
- The experiments, which were conducted on two real world datasets, support the viability of our approach.

References

[1] B. D. Lucas, T. Kanade et al., “An iterative image registration technique with an application to stereo vision,” 1981.

[2] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan, “Fourier lucas-kanade algorithm,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 6, pp. 1383–1396, 2013.

[3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, “Orb:An efficient alternative to sift or surf.” in ICCV, vol. 11, no. 1.Citeseer, 2011, p. 2.

[5] K. Takeda, K. Kise, and M. Iwamura, “Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah,” in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 1054–1058.

[6] M. Block, M. R. Ortegon, A. Seibert, J. Kretzschmar, and R. Rojas, “Sitt-a simple robust scaleinvariant text feature detector for document mosaicing,” Proc. of ICIA2009, pp. 400–403, 2007.

[7] E. Royer, J. Chazalon, M. Rusinol, and F. Bouchara, “Benchmarking keypoint filtering approaches for document image matching,” in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, 2017, pp. 343–348.

[8] I. Rocco, R. Arandjelovic, and J. Sivic, “Convolutional neural network architecture for geometric matching,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6148–6157.

[9] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Deep image homography estimation,” arXiv preprint arXiv:1606.03798, 2016.

[10] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, “Unsupervised deep homography: A fast and robust homography estimation model,” IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 2346–2353, 2018.

[11] Tian, Zhi, et al. "Detecting text in natural image with connectionist text proposal network." *European conference on computer vision*. Springer, Cham, 2016.

[12] U.-V. Marti and H. Bunke, “The iam-database: an english sentence database for offline handwriting recognition,” International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 39–46, 2002.

[13] Chowdhury, Arindam, and Lovekesh Vig. "An efficient end-to-end neural model for handwritten text recognition." arXiv preprint arXiv:1807.07965 (2018).

Thank You!

Contact us:

TCS Research, New Delhi
India

kushagra.mahajan@tcs.com